en by

$$\hat{H}_i(t) = \exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)\hat{H}_0(t). \tag{3.21}$$

ı multiplying each side of equation (3.20) by -1 and exponentiating, and ıking use of equation (1.5), we find that the estimated survivor function for e ith individual is

$$\hat{S}_i(t) = \left\{\hat{S}_0(t)\right\}^{\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)}, \tag{3.22}$$

$t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$. Note that once the estimated sur- vor function, $\hat{S}_i(t)$, has been obtained, an estimate of the cumulative hazard nction is simply $-\log \hat{S}_i(t)$.

8.1 The special case of no covariates

hen there are no covariates, so that we have just a single sample of survival mes, equation (3.16) becomes

$$\frac{d_j}{1-\hat{\xi}_j} = n_j,$$

om which

$$\hat{\xi}_j = \frac{n_j - d_j}{n_j}.$$

hen, the estimated baseline hazard function at time $t_{(j)}$ is $1-\hat{\xi}_j$, which ; d_j/n_j. The corresponding estimate of the survivor function from equa- ion (3.18) is $\prod_{j=1}^{k}\hat{\xi}_j$, that is,

$$\prod_{j=1}^{k}\left(\frac{n_j - d_j}{n_j}\right),$$

vhich is the Kaplan-Meier estimate of the survivor function given earlier in equation (2.4). This shows that the estimate of the survivor function given in equation (3.22) generalises the Kaplan-Meier estimate to the case where the ıazard function depends on explanatory variables.

Furthermore, the estimate of the hazard function in equation (3.17) reduces to $d_j/\{n_j(t_{(j+1)} - t_{(j)})\}$, which is the estimate of the hazard function given in equation (2.16) of Chapter 2.

3.8.2 Some approximations to estimates of the baseline functions

When there are tied survival times, the estimated baseline hazard can only be found by using an iterative method to solve equation (3.16). This iterative process can be avoided by using an approximation to the summation on the left-hand side of equation (3.16).

The term

$$\hat{\xi}_j^{\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l)},$$

in the denominator of the left-hand side of equation (3.16), can be written as

$$\exp\left\{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l}\log\hat{\xi}_j\right\},$$

and taking the first two terms in the expansion of the exponent gives

$$\exp\left\{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l}\log\hat{\xi}_j\right\} \approx 1 + e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l}\log\hat{\xi}_j.$$

Writing $1-\tilde{\xi}_j$ for the estimated baseline hazard at time $t_{(j)}$, obtained using this approximation, and substituting $1 + e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l}\log\tilde{\xi}_j$ for $\hat{\xi}_j^{\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l)}$ in equation (3.16), we find that $\tilde{\xi}_j$ is such that

$$-\sum_{l\in D(t_{(j)})}\frac{1}{\log\tilde{\xi}_j} = \sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l).$$

Therefore,

$$\frac{-d_j}{\log\tilde{\xi}_j} = \sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l),$$

since d_j is the number of deaths at the jth ordered death time, $t_{(j)}$, and so

$$\tilde{\xi}_j = \exp\left(\frac{-d_j}{\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l)}\right). \tag{3.23}$$

From equation (3.18), an estimate of the survivor function, based on the values of $\tilde{\xi}_j$, is given by

$$\tilde{S}_0(t) = \prod_{j=1}^{k}\exp\left(\frac{-d_j}{\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l)}\right), \tag{3.24}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$. From this definition, the estimated survivor function is not necessarily zero at the longest survival time, when that time is uncensored, unlike the estimate in equation (3.18). The estimate of the baseline cumulative hazard function derived from $\tilde{S}_0(t)$ is

$$\tilde{H}_0(t) = -\log\tilde{S}_0(t) = \sum_{j=1}^{k}\frac{d_j}{\sum_{l\in R(t_{(j)})}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l)}, \tag{3.25}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$. This estimate is often referred to as the *Nelson-Aalen estimate* or the *Breslow estimate* of the baseline cumulative hazard function.

When there are no covariates, the estimated baseline survivor function in equation (3.24) becomes

$$\prod_{j=1}^{k}\exp(-d_j/n_j), \tag{3.26}$$

since n_j is the number of individuals at risk at time $t_{(j)}$. This is the Nelson-Aalen estimate of the survivor function given in equation (2.6) of Chapter 2, and the corresponding estimate of the baseline cumulative hazard function is $\sum_{j=1}^{k} d_j/n_j$, as in Section 2.3.3 of Chapter 2.

A further approximation is found from noting that the expression

$$\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)},$$

in the exponent of equation (3.23), will tend to be small, unless there are large numbers of ties at particular death times. Taking the first two terms of the expansion of this exponent, and denoting this new approximation to ξ_j by ξ_j^* gives

$$\xi_j^* = 1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}.$$

Adapting equation (3.17), the estimated baseline hazard function in the interval from $t_{(j)}$ to $t_{(j+1)}$ is then given by

$$h_0^*(t) = \frac{d_j}{(t_{(j+1)} - t_{(j)}) \sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}, \tag{3.27}$$

for $t_{(j)} \leqslant t < t_{(j+1)}$, $j = 1, 2, \ldots, r-1$. Using ξ_j^* in place of $\hat{\xi}_j$ in equation (3.18), the corresponding estimated baseline survivor function is

$$S_0^*(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)} \right),$$

and a further approximate estimate of the baseline cumulative hazard function is $H_0^*(t) = -\log S_0^*(t)$. Notice that the cumulative hazard function in equation (3.25) at time t can be expressed in the form

$$\tilde{H}_0(t) = \sum_{j=1}^{k} (t_{(j+1)} - t_{(j)}) h_0^*(t),$$

where $h_0^*(t)$ is given in equation (3.27). Consequently, differences in successive values of the estimated baseline cumulative hazard function in equation (3.25) provide an approximation to the baseline hazard function, at times $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, that can easily be computed.

In the particular case where there are no covariates, the estimates $h_0^*(t)$, $S_0^*(t)$ and $H_0^*(t)$ are the same as those given in Section 3.8.1. Equations similar to equations (3.21) and (3.22) can be used to estimate the cumulative hazard and survivor functions for an individual whose vector of explanatory variables is $\boldsymbol{x}_i$.

In practice, it will often be computationally advantageous to use either $\tilde{S}_0(t)$ or $S_0^*(t)$ in place of $\hat{S}_0(t)$. When the number of tied survival times is small, all three estimates will tend to be very similar. Moreover, since the estimates are

generally used as descriptive summaries of the survival data, small differences between the estimates are unlikely to be of practical importance.

Once an estimate of the survivor function has been obtained, the median and other percentiles of the survival time distribution can be found from tabular or graphical displays of the function for individuals with particular values of explanatory variables. The method used is very similar to that described in Section 2.4, and is illustrated in the following example.

Example 3.12 Treatment of hypernephroma

In Example 3.4, a proportional hazards model was fitted to the data on the survival times of patients with hypernephroma. The hazard function was found to depend on the age group of a patient, and whether or not a nephrectomy had been performed. The estimated hazard function for the ith patient was found to be

$$\hat{h}_i(t) = \exp\{0.013\, A_{2i} + 1.342\, A_{3i} - 1.412\, N_i\} \hat{h}_0(t),$$

where A_{2i} is unity if the patient is aged between 60 and 70 and zero otherwise, A_{3i} is unity if the patient is aged over 70 and zero otherwise, and N_i is unity if the patient has had a nephrectomy and zero otherwise. The estimated baseline hazard function is therefore the estimated hazard of death at time t, for an individual whose age is less than 60 and who has not had a nephrectomy.

In Table 3.10, the estimated baseline hazard function, $\hat{h}_0(t)$, cumulative hazard function, $\hat{H}_0(t)$, and survivor function, $\hat{S}_0(t)$, obtained using equations (3.15), (3.19) and (3.18), respectively, are tabulated.

From this table, we see that the general trend is for the estimated baseline hazard function to increase with time. From the manner in which the estimated baseline hazard function has been computed, the estimates only apply at the death times of the patients in the study. However, if the assumption of a constant hazard in each time interval is made, by dividing the estimated hazard by the corresponding time interval, the risk of death per unit time can be found. This leads to the estimate in equation (3.17). A graph of this hazard function is shown in Figure 3.2.

This graph shows that the risk of death per unit time is roughly constant over the duration of the study. Table 3.10, also shows that the values of $\hat{h}_0(t)$ are very similar to differences in the values of $\hat{H}_0(t)$ between successive observations, as would be expected.

We now consider the estimation of the median survival time, which is the smallest observed survival time for which the estimated survivor function is less than 0.5. From Table 3.10, the estimated median survival time for patients aged less than 60 who have not had a nephrectomy is 12 months.

By raising the estimate of the baseline survivor function to a suitable power, the estimated survivor functions for patients in other age groups, and for patients who have had a nephrectomy, can be obtained through equation (3.22). Thus, the estimated survivor function for the ith individual is given by

$$\hat{S}_i(t) = \left\{ \hat{S}_0(t) \right\}^{\exp\{0.013 A_{2i} + 1.342 A_{3i} - 1.412 N_i\}}.$$

Table 3.10 *Estimates of the baseline hazard and survivor functions for the data from Example 3.4.*

Time	$\hat{h}_0(t)$	$\hat{S}_0(t)$	$\hat{H}_0(t)$
0	0.000	1.000	0.000
5	0.050	0.950	0.051
6	0.104	0.852	0.161
8	0.113	0.755	0.281
9	0.237	0.576	0.552
10	0.073	0.534	0.628
12	0.090	0.486	0.722
14	0.108	0.433	0.836
15	0.116	0.383	0.960
17	0.132	0.333	1.101
18	0.285	0.238	1.436
21	0.185	0.194	1.641
26	0.382	0.120	2.123
35	0.232	0.092	2.387
36	0.443	0.051	2.972
38	0.279	0.037	3.299
48	0.299	0.026	3.655
52	0.560	0.011	4.476
56	0.382	0.007	4.958
68	0.421	0.004	5.504
72	0.467	0.002	6.134
84	0.599	0.001	7.045
108	0.805	0.000	8.692
115	–	0.000	–

For an individual aged less than 60 who has had a nephrectomy, $A_2 = 0$, $A_3 = 0$, and $N = 1$, so that the estimated survivor function for this individual is

$$\left\{\hat{S}_0(t)\right\}^{\exp\{-1.412\}}.$$

This function is plotted in Figure 3.3, together with the estimated baseline survivor function, which is for an individual in the same age group but who has not had a nephrectomy.

This figure shows that the probability of surviving beyond any given time is greater for those who have had a nephrectomy, confirming that a nephrectomy improves the prognosis for patients with hypernephroma.

Note that because of the assumption of proportional hazards, the two estimated survivor functions in Figure 3.3 cannot cross. Moreover, the estimated survivor function, for those who have had a nephrectomy, lies above that of those on whom a nephrectomy has not been performed. This is a direct con-

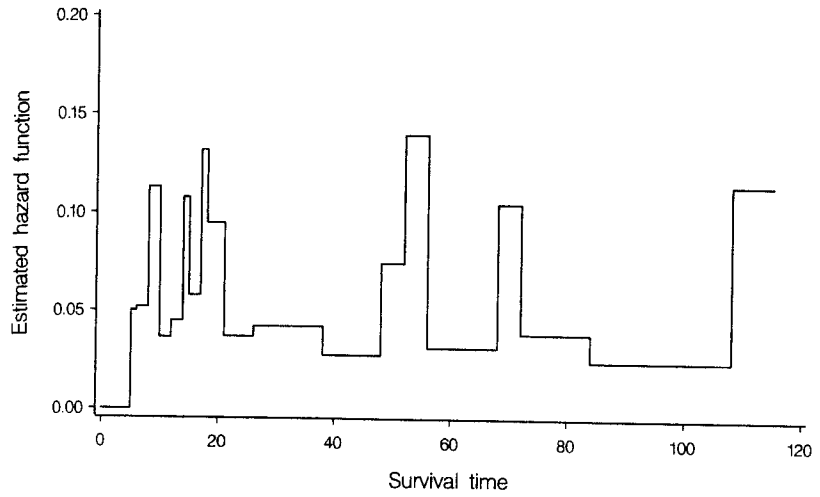

Figure 3.2 *Estimated baseline hazard function, per unit time, assuming constant hazard between adjacent death times.*

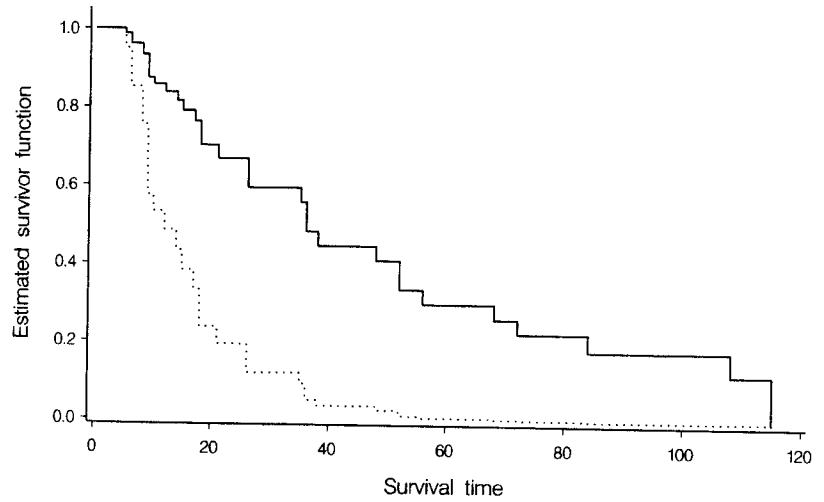

Figure 3.3 *Estimated survivor functions for patients aged less than 60, with (—) and without (· · ·) a nephrectomy.*

sequence of the estimated hazard ratio for those who have had the operation, relative to those who have not, being less than unity.

An estimate of the median survival time for this type of patient can be obtained from the tabulated values of the estimated survivor function, or from the graph in Figure 3.3. We find that the estimated median survival time for a patient aged less than 60 who has had a nephrectomy is 36 months. Other percentiles of the distribution of survival times can be estimated using a similar approach.

In a similar manner, the survivor functions for patients in the different age groups can be compared, either for those who have had or not had a nephrectomy. For example, for patients who have had a nephrectomy, the estimated survivor functions for patients in the three age groups are respectively $\{\hat{S}_0(t)\}^{\exp\{-1.412\}}$, $\{\hat{S}_0(t)\}^{\exp\{-1.412+0.013\}}$ and $\{\hat{S}_0(t)\}^{\exp\{-1.412+1.342\}}$. These estimated survivor functions are shown in Figure 3.4.

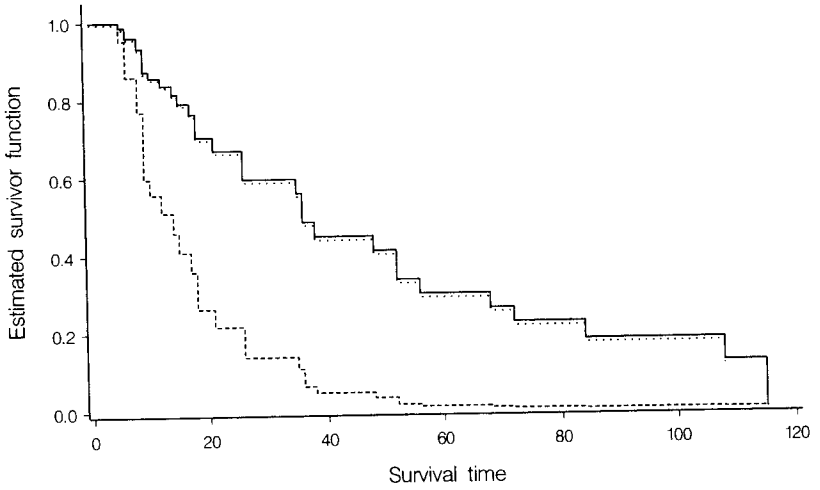

Figure 3.4 *Estimated survivor functions for patients aged less than 60 (—), between 60 and 70 (···) and greater than 70 (- - -), who have had a nephrectomy.*

This figure clearly shows that patients aged over 70 have a poorer prognosis than those in the other two age groups.

3.9* Proportional hazards modelling and the log-rank test

The proportional hazards model can be used to test the null hypothesis that there is no difference between the hazard functions for two groups of survival times, as illustrated in Example 3.3. This modelling approach therefore provides an alternative to the log-rank test in this situation. However, there is a close connection between the two procedures, which is explored in greater detail in this section.

Following the notation used in Section 2.6.2, and summarised in Table 2.7, the two groups will be labelled Group I and Group II, respectively. The numbers of individuals in the two groups who die at the jth ordered death time, $t_{(j)}$, $j = 1, 2, \ldots, r$, will be denoted by d_{1j} and d_{2j}, respectively. Similarly, the numbers of individuals at risk in the two groups at time $t_{(j)}$, that is, the numbers who are alive and uncensored just prior to this time, will be denoted n_{1j} and n_{2j}, respectively.

Now let X be an indicator variable that is unity when an individual is in Group I and zero when an individual is in Group II. The proportional hazards model for the ith individual can be written as

$$h_i(t) = e^{\beta x_i} h_0(t),$$

where x_i is the value of X for the ith individual, $i = 1, 2, \ldots, n$. When there are no tied observations, that is, when $d_j = d_{1j} + d_{2j} = 1$, this model can be fitted by finding that value $\hat{\beta}$ which maximises the likelihood function in equation (3.4). Denoting the value of X for the individual who dies at $t_{(j)}$ by $x_{(j)}$, the likelihood function is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta x_{(j)})}{\sum_{l=1}^{n_j} \exp(\beta x_l)}, \tag{3.28}$$

since there are $n_j = n_{1j} + n_{2j}$ individuals in the risk set, $R(t_{(j)})$, at time $t_{(j)}$, and the corresponding log-likelihood function is

$$\log L(\beta) = \sum_{j=1}^{r} \beta x_{(j)} - \sum_{j=1}^{r} \log \left\{ \sum_{l=1}^{n_j} \exp(\beta x_l) \right\}.$$

Since $x_{(j)}$ is zero for individuals in Group II, the first summation in this expression is over the death times in Group I, and so is simply $d_1\beta$, where $d_1 = \sum_{j=1}^{r} d_{1j}$ is the total number of deaths in Group I. Also,

$$\sum_{l=1}^{n_j} \exp(\beta x_l) = n_{1j} e^{\beta} + n_{2j},$$

and so

$$\log L(\beta) = d_1\beta - \sum_{j=1}^{r} \log \left\{ n_{1j} e^{\beta} + n_{2j} \right\}. \tag{3.29}$$

The maximum likelihood estimate of β can be found by maximising this expression with respect to β, for which a non-linear optimisation routine is required. Then, the null hypotheses that $\beta = 0$ can be tested by comparing the value of $-2 \log \hat{L}(\hat{\beta})$ with $-2 \log \hat{L}(0)$. This latter quantity is simply $2 \sum_{j=1}^{r} \log n_j$.

Computation of $\hat{\beta}$ can be avoided by using a *score test* of the null hypothesis that $\beta = 0$. This test procedure, which is outlined in Appendix A, is based on the test statistic

$$\frac{u^2(0)}{i(0)},$$

where

$$u(\beta) = \frac{\partial \log L(\beta)}{\partial \beta}$$

is the efficient score, and

$$i(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta^2}$$

is Fisher's (observed) information function. Under the null hypothesis that $\beta = 0$, $u^2(0)/i(0)$ has a chi-squared distribution on one degree of freedom.

Now, from equation (3.29),

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{j=1}^{r} \left(d_{1j} - \frac{n_{1j}e^{\beta}}{n_{1j}e^{\beta} + n_{2j}} \right),$$

and

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial \beta^2} &= -\sum_{j=1}^{r} \frac{(n_{1j}e^{\beta} + n_{2j})n_{1j}e^{\beta} - (n_{1j}e^{\beta})^2}{(n_{1j}e^{\beta} + n_{2j})^2} \\ &= -\sum_{j=1}^{r} \frac{n_{1j}n_{2j}e^{\beta}}{(n_{1j}e^{\beta} + n_{2j})^2}. \end{aligned}$$

The efficient score and information function, evaluated at $\beta = 0$, are therefore given by

$$u(0) = \sum_{j=1}^{r} \left(d_{1j} - \frac{n_{1j}}{n_{1j} + n_{2j}} \right),$$

and

$$i(0) = \sum_{j=1}^{r} \frac{n_{1j}n_{2j}}{(n_{1j} + n_{2j})^2}.$$

These are simply the expressions for U_L and V_L given in equations (2.23) and (2.25) of Chapter 2, for the special case where there are no ties, that is, where $d_j = 1$ for $j = 1, 2, \ldots, r$.

When there are tied observations, the likelihood function in equation (3.28) has to be replaced by one that allows for ties. In particular, if the likelihood function in equation (3.11) is used, the efficient score and information function are exactly those given in equations (2.23) and (2.25). Hence, when there are tied survival times, the log-rank test corresponds to using the score test for the discrete proportional hazards model due to Cox (1972). In practice, the P-value that results from this score test will not usually differ much from that obtained from comparing the values of the statistic $-2 \log \hat{L}$ for the models with and without a term corresponding to the treatment effect. This was noted in the discussion of Example 3.3. Of course, one advantage of using the Cox regression model in the analysis of such data is that it leads directly to an estimate of the hazard ratio.

3.10 Further reading

Comprehensive introductions to statistical modelling in the context of linear regression analysis are given by Draper and Smith (1981) and Montgomery *et al.* (2001). McCullagh and Nelder (1989) include a chapter on models for survival data in their encyclopaedic survey of generalised linear modelling. Aitkin *et al.* (1989) illustrate the theory and practice of linear modelling through the statistical package GLIM, and also include a chapter on the analysis of survival data.

Model formulation and strategies for model selection are discussed in books on linear regression analysis, and also in Chapter 5 of Chatfield (1995), Chapter 4 of Cox and Snell (1981), and Appendix 2 of Cox and Snell (1989). Miller (2002) describes a wide range of procedures for identifying suitable subsets of variables to use in linear regression modelling. What has come to be known as Akaike's information criterion was introduced by Akaike (1974). It is widely used in times series analysis and described in books on this subject, such as Chatfield (1996) and Janacek (2001). The hierarchic principle is fully discussed by Nelder (1977), and in Chapter 3 of McCullagh and Nelder (1989). Harrell (2001) addresses many practical issues in model building and illustrates the process using two extensive case studies involving survival data.

The proportional hazards model for survival data, in which the baseline hazard function remains unspecified, was proposed by Cox (1972). This paper introduced the notion of partial likelihood, which was subsequently considered in greater detail by Cox (1975). See also the contributions to the discussion of Cox (1972) by Kalbfleisch and Prentice (1972) and Breslow (1972). A detailed review of the model, and extensions of it, is contained in Therneau and Grambsch (2000).

Introductions to the proportional hazards model, intended for medical researchers have been given by Christensen (1987), Elashoff (1983) and Tibshirani (1982). More recent accounts are given in the textbooks referenced in Section 1.4 of Chapter 1. In particular, Hosmer and Lemeshow (1999) include a careful discussion on model development and the interpretation of model-based parameter estimates.

A detailed treatment of ties in survival data is given in Kalbfleisch and Prentice (2002) and Lawless (2002); see also Breslow (1972) and Peto (1972). DeLong *et al.* (1994) give an equivalent expression for the exact partial likelihood in the presence of ties that has computational advantages. The estimate of the baseline survivor function, denoted by $\hat{S}_0(t)$ in Section 3.8, was introduced by Kalbfleisch and Prentice (1973) and is also described in Kalbfleisch and Prentice (2002). The estimate $S_0^*(t)$ was presented by Breslow (1972, 1974), although it was derived using a different argument from that used in Section 3.8.2.